Texte paru dans le n°2 de la revue Corpus, 2003, p. 95-118.

La distance intertextuelle[1]

Cyril LABBE*, Dominique LABBE**
*IMAG, Université Joseph Fourrier, Grenoble
**Institut d'Etudes Politiques de Grenoble

Résumé : Présentation de l'indice de la distance intertextuelle et de ses propriétés. Discussion des limites du calcul : influence des décimales et des différences de taille entre textes comparés. Examen de la contribution à la distance des vocables classés en fonction de leurs catégories grammaticales et de leurs fréquences. L'indice de la distance intertextuelle fournit un outil intéressant pour la mesure des ressemblances et des dissemblances au sein des grandes bases de textes.

Abstract : This paper proposes a presentation of the calculation of intertextual distance index and its properties. The limitations of the calculation formula are a subject of discussion; i.e., the decimal influence and the difference in size between the compared texts. An examination of the contribution to the measurement of the distance by the organization of the vocables according to their grammatical categories and to their frequency is also performed. The intertextual distance provides a robust and reliable tool for measuring similarities as well as dissimilarities within large text bases.

Comment mesurer la ressemblance, ou la dissemblance entre deux textes ? Une réponse précise à cette question permettrait de résoudre de nombreux problèmes posés à la statistique lexicale. L'"attribution d'auteur" constitue certainement le plus connu de ces problèmes : à quelles conditions peut-on désigner, avec un degré raisonnable de certitude, l'auteur d'un texte d'origine douteuse ou inconnue et différencier ainsi cette "plume de l'ombre" de toutes les autres possibles ? Parmi les nombreuses réponses possibles, la plus prometteuse semble porter sur la mesure de la distance existant entre les vocabulaires des

[1] Les auteurs remercient X. Luong et T. Merriam qui ont relu avec attention une première version de ce travail et dont les remarques ont permis d'améliorer sensiblement notre texte.

textes soumis à comparaison (voir à ce sujet : Holmes, 1995 ; Baayen et al, 1996 ; Rudman, 1998 ; Labbé et Labbé, 2001).

Cette discussion débouche sur une interrogation plus fondamentale et plus stratégique : dans les vastes ensembles de textes électroniques aujourd'hui à la disposition du chercheur, il s'agit de constituer, de manière automatique, des groupes plus ou moins homogènes du point de vue de leurs vocabulaires, genres et thèmes. De tels outils seraient précieux pour la critique littéraire, la lexicologie, la recherche documentaire, la traduction automatique…

La question est connue sous le nom de «**connexion** lexicale». Celle-ci est définie comme «l'intersection du vocabulaire de deux textes» (Muller 1977, p 145-154 ; Brunet, 1988a). La connexion est donc le complémentaire de la **distance**. Nous avons retenu ce dernier terme car il est bien connu en statistique (Gower, 1985).

Pour comprendre la portée de ce calcul, il faut rappeler la différence existant entre « **mot** » et « mot différent » (ou **vocable**). Le mot est le plus petit élément mesurable d'un texte et le vocable, forme l'élément de base du **vocabulaire**. Par exemple, le plus long roman en langue française, *Les misérables,* compte près d'un demi-million de mots (c'est sa **taille** ou son « étendue », notée N) et son vocabulaire (noté V) comporte moins de 12.000 vocables.

Jusqu'à maintenant, l'étude de la «connexion» a été faite sur le vocabulaire sans tenir compte de la fréquence des vocables. Ces calculs étaient inspirés des sciences de la vie et habituellement connus sous le nom d'indices «de Jaccard» (Hubalek, 1982). On relève simplement la présence ou l'absence des individus dans les populations comparées sans tenir compte de leurs effectifs : peu importe qu'un vocable soit employé une fois ou mille fois, il compte toujours pour 1 dans le calcul. En zoologie, en biologie ou en épidémiologie, de telles techniques s'imposaient puisqu'il a été longtemps impossible, ou peu utile, d'estimer la densité des individus observés. Cette impossibilité explique que, aujourd'hui encore, la plupart des formules utilisées dans les sciences du vivant négligent les fréquences pour ne retenir que les seules présence ou absence des individus. Il n'y a aucune raison pour agir de même en statistique lexicale puisque, avec l'aide de l'ordinateur, la fréquence d'emploi de chaque vocable dans un

texte donné est connue avec une totale précision. Considérer la fréquence d'emploi de chacun des vocables, c'est prendre en compte la surface totale des textes comparés (il est vrai que l'on déstructure le texte, mais la discussion de ce problème nous entraînerait trop loin). Dans le terme «distance intertextuelle», l'adjectif **textuel** indique donc que les calculs portent sur l'ensemble des textes (N) et non sur leur seul vocabulaire (V).

Après avoir présenté le calcul nous discuterons des limites du procédé et du degré de confiance que l'on peut accorder aux résultats. Enfin, nous montrerons que tous les vocables n'interviennent pas de la même manière dans cette distance, que certains vocables, rares ou appartenant à des catégories grammaticales précises, ont un rôle prépondérant, mais souvent mal connu, dans la «fabrication du sens».

1. Le calcul de l'indice de la distance intertextuelle

Pour pouvoir dire si des textes sont «plutôt proches» ou «plutôt éloignés», quant à l'utilisation des mots, on cherche à exprimer par un nombre, des propriétés – la ressemblance ou la dissemblance — qui ne sont pas des nombres. De plus, pour pouvoir répéter cette mesure autant de fois qu'on voudra, il faudra encore transformer la mesure absolue en un indice. Pour remplir ces deux objectifs, il faudra que l'indice de la distance (noté D dans la suite de cet article) présente les propriétés suivantes :

— insensible aux différences de taille entre les textes ;

— applicable à plusieurs textes et, potentiellement, à tous les textes d'une même langue ;

— variant uniformément — entre 0 (même vocabulaire et fréquence semblable de chacun des vocables dans les deux textes) et 1 (aucun vocable en commun) — sans saut ni effet de seuil autour de certaines valeurs ;

— symétrique (soit deux textes A et B alors $D(A,B) = D(B,A)$) ;

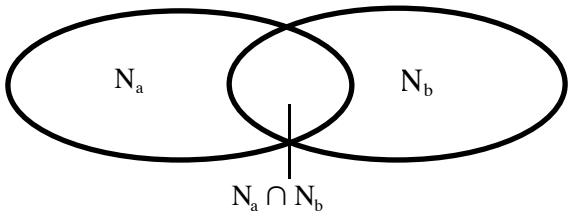
— vérifiant l'inégalité triangulaire : $D(A,B) \leq D(A,C) + D(C,B)$. Ce qui revient à dire que le chemin entre deux individus sera toujours plus long si l'on passe par un troisième (à moins que ce troisième soit confondu avec l'un des deux premiers) ;

— enfin l'indice doit être "transitif". Quand on agrège le vocabulaire de deux textes, les indices des distances de ce nouveau «corpus» vis-à-vis des autres textes doivent refléter l'ordre des indices antérieurs : si D(A,B) > D(A,C) > D(B,C) alors : D(A,B) > D[A,(B∪C)] ;

— aussi «robuste» que possible (ie une modification marginale dans le vocabulaire d'un des deux textes doit se traduire par une variation marginale de l'indice)…

1.1 Connexion lexicale et distance intertextuelle

Quand on examine les travaux classiques en ce domaine, on trouve habituellement le raisonnement suivant. Soit deux textes A et B. La distance absolue entre ces deux textes sera la réunion des deux textes (A ∪ B) moins leur intersection (A ∩ B) :



On trouve dans la littérature les formules :

$$(1)\quad D_{(a,b)} = \frac{\sum_{V \in A} | F_{ia} - F_{ib} | + \sum_{V \in B} | F_{ib} - F_{ia} |}{N_a + N_b}$$

ou :

$$(2)\quad D_{(a,b)} = \frac{1}{2} \left(\frac{\sum_{V \in A} | F_{ia} - F_{ib} |}{N_a} + \frac{\sum_{V \in B} | F_{ib} - F_{ia} |}{N_b} \right)$$

Avec :
V_a et V_b : nombre de vocables dans A et B ;
F_{ia} : fréquence du vocable i dans A ;
F_{ib} : fréquence du vocable i dans B.
N_a et N_b : nombre de mots dans A et B (taille) ;

et : $N_a = \sum F_{ia}$ et $N_b = \sum F_{ib}$

La formule (1) est dans la lignée de la "connexion lexicale" introduite par C. Muller (Muller 1977). Ces formules soulèvent deux remarques :

— (1) et (2) ne sont équivalentes que quand les textes ont des tailles égales ($N_a = N_b$). Si les deux textes comparés ne partagent aucun vocable, les formules (1) et (2) donnent bien un indice de 1 quelle que soit la taille des textes (ce qui est une des conditions requises pour l'indice idéal). En revanche, le minimum théorique n'atteint zéro que dans le cas particulier de tailles égales. En effet, plus les textes comparés seront de tailles différentes, plus le numérateur minimal possible s'éloignera de zéro. Par exemple, dans le premier corpus auquel nous avons appliqué ce calcul — les discours du trône des gouvernements québécois devant les chambres provinciales (Labbé et Monière, 2000) — le discours de 1965, qui est le plus court du corpus, a une taille de 1 006 mots et un vocabulaire de 419 vocables alors que le texte de 1984 (le plus long du corpus) contient 12 828 mots et 2 790 vocables. Physiquement parlant, les 2 790 vocables du texte de 1984 ne peuvent pas tous entrer dans le texte de 1965. Même si le petit texte était totalement inclus dans le grand, la distance ne serait pas nulle, puisque le calcul porte également sur les (2 790 - 419) vocables absents du plus petit et ne pouvant pas tous y figurer. Au contraire, dans l'hypothèse la plus favorable — le texte de 1965 est un «modèle réduit» de celui de 1984 —, l'indice de la distance des vocabulaires (ne tenant pas compte des fréquences) serait encore de 15,2% : (2790-419)/(12828+2790).

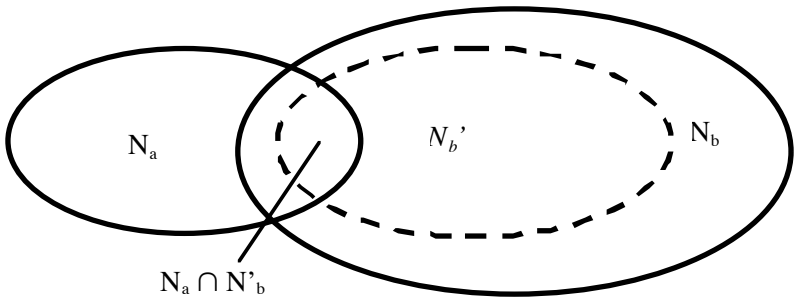
— dans (1) comme dans (2), les vocables communs aux deux textes interviennent deux fois dans le calcul. On leur donne donc plus d'importance qu'aux vocables propres. Or, ce vocabulaire commun est d'autant plus faible que les tailles des textes sont différentes.

En fait, ces formules mesurent moins la distance entre les textes que leur différence de taille. Elles conduisent nécessairement à classer ensemble les textes courts (poésie), moyens (théâtre) et longs (roman) (Voir par exemple : Brunet, 1978).

Nous proposons de surmonter ces objections de la manière suivante.

1.2 Une approximation de la distance intertextuelle

Il est proposé de simuler la réduction du plus grand des deux textes à la taille du plus petit. Soit B' cette réduction de B en fonction de la taille de A :



Avec :

$$U_{(a,b)} = \frac{N_a}{N_b}$$

Tout vocable de fréquence F_i dans B aura une fréquence attendue dans A égale à :

$$E_{ia(u)} = F_{ib} * U_{(a,b)}$$

D'où l'on tire la taille du texte B réduit :

$$N'_b = \sum_{V \in B'} E_{ia(u)}$$

Dans la formule (1), on remplace les termes F_{ib} par $E_{ia(u)}$ et N_b par N'_b. Le minimum théorique (zéro) sera atteint quand le petit texte sera une sorte de modèle réduit du grand.

Dans ce cas, tous les vocables de A se retrouvent dans B avec une fréquence telle que :

$$F_{ia} = E_{ia(u)}$$

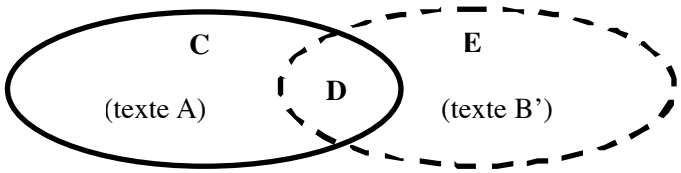
Le numérateur sera égal à zéro et le dénominateur à :

$$N_a + N'_b = 2\ N_a$$

C'est en effet l'effectif maximum des mots que les deux textes peuvent partager s'ils ont même dimension, même vocabulaire et, pour chacun des vocables, même fréquence. Le maximum théorique (l'unité) devrait être atteint quand les deux textes n'ont aucun mot en commun. Au numérateur, comme au dénominateur, figureront N_a et N'_b.

Toutefois, cette nouvelle formulation ne répond pas à l'objection concernant le double compte des mots communs aux deux textes et elle ne résout pas totalement le problème physique mentionné ci-dessus : tous les vocables de B ne peuvent pas théoriquement figurer dans A. Pour tenir compte de ces deux objections, il est proposé de procéder de la manière suivante (figure ci-dessous) :

- ne considérer qu'une seule fois les mots communs aux deux textes ;
- limiter le calcul à l'ensemble des vocables de A mais aux seuls vocables de B dont la fréquence est telle que l'on en attend au moins 1 dans A ($E_{ia(u)} \geq 1$). La somme de ces fréquences théoriques donne N'_b.



La distance absolue entre A et B' sera :

$$D_{Va,b(u)} = \sum_{V\in A,\ V\in B'} \left| F_{ia} - E_{ia(u)} \right|$$

Quant A et B n'ont aucun mot en commun, la distance entre eux sera égale à $N_a + N'_b$. Cette somme sera naturellement placée au dénominateur de la formule de l'indice de la distance (3) ci-dessous : ainsi la valeur maximale sera égale à 1 et, l'indice sera nécessairement inférieur à 1 quand l'intersection des deux textes ne sera pas vide (ce qui sera toujours le cas lorsqu'ils sont écrits dans la même langue).

$$(3)\quad D_{(a,b)} = \frac{\sum_{V\in A,\ V\in B} \left| F_{ia} - E_{ia(u)} \right|}{\sum_{V\in A} F_{ia} + \sum_{V\in B} E_{ia(u)}} = \frac{\sum_{V\in A,\ V\in B} \left| F_{ia} - E_{ia(u)} \right|}{N_a + N'_b}$$

On remarquera que le même résultat, aux arrondis près, peut être obtenu en soustrayant les fréquences relatives de chacun des vocables dans les deux textes comparés, à condition de limiter le calcul à tout le vocabulaire du plus petit des deux textes et à ceux des vocables qui, dans le plus grand, ont une fréquence suffisante pour qu'on en attende au moins un s'il avait la taille du plus petit.

Depuis plusieurs années, nous avons appliqué ce calcul à un grand nombre de corpus comportant plusieurs milliers de textes en français moderne, et plus de dix millions de mots, tous dépouillés selon la même norme (Labbé, 1990). Citons notamment : les discours du général de Gaulle et de F. Mitterrand, les discours du trône des Premiers ministres canadiens et québécois depuis 1945 (Labbé-Monière, 2000), des éditoriaux de la presse syndicale (Labbé-Brugidou, 1999), des articles de la presse économique, des transcriptions d'entretiens sociologiques (Bergeron-Labbé, 2000).

Mais il ne suffit pas de dire : «ça marche !», encore faut-il valider la méthode, notamment par des expériences bien conçues. L'une de ces expériences a été réalisée "en double aveugle" grâce à E. Brunet : celui-ci nous a envoyé 50 textes anonymés — de tailles comprises entre 8 100 et 9 200 mots —, à charge pour nous d'identifier ceux tirés d'un même livre ou écrits par un même auteur dans des ouvrages différents. En fait, il y avait 11 auteurs

différents, avec chacun 2 livres et deux extraits par livres. Enfin, E. Brunet avait placé 6 chimères créées en collant des passages extraits dans chacun des textes[2]. Nous allons utiliser ces textes ("corpus Brunet" dans la suite de cet article) pour mesurer dans quelles limites s'inscrivent ces réussites et quel degré de confiance accorder aux résultats.

2. Les limites du calcul

En premier lieu, on aura remarqué que la distance intertextuelle n'est pas calculée sur l'ensemble des textes et que ce défaut sera d'autant plus net que les différences de taille entre les deux textes seront importantes. De plus, les écarts entre les valeurs observées et les valeurs attendues ne sont pas élevés au carré comme le font les analyses en composantes principales ou factorielles[3] (d'autres méthodes de classification, abordées par ailleurs dans ce numéro, permettent de pallier cette absence de «visualisation»). Deux autres problèmes méritent d'être signalés : les arrondis introduisent une certaine marge d'incertitude dans le calcul ; de plus, l'indice n'est pas tout à fait insensible aux différences de taille entre les textes.

2.1 L'influence des décimales

Le calcul consiste à soustraire des entiers — les fréquences observées — à des nombres rationnels (les fréquences théoriques) qui comporteront quasiment toujours des décimales (une fraction de mot !), ce «reste» entrant dans la distance et interdisant qu'elle puisse être nulle (sauf lorsque les deux textes ont le même nombre de mots ou lorsque la taille du plus grand est un multiple de celle du petit). Pour évaluer l'incidence de ce biais, nous avons relevé ces «fractions de mots» (la partie décimale de la

[2] Un compte rendu de cette expérience peut être consulté sur la page personnelle de D. Labbé (http://www.upmf-grenoble.fr/cerat/).

[3] En fait, ces techniques sont probablement peu adaptées aux populations qui, comme les phénomènes langagiers, sont composées d'un très grand nombre d'individus dont les occurrences sont peu nombreuses et, surtout, très inégalement réparties. Cette discussion dépasserait l'objet de cet article.

distance), au cours du calcul des 2 352 distances entre les 50 textes du corpus Brunet pris deux à deux.

La comparaison porte sur un effectif théorique total de 39 641 366,07 mots et génère une distance absolue de 15 002 457,87 mots. Nous avons soustrait de la distance, calculée sur chacun des vocables, l'entier immédiatement supérieur lorsque la partie décimale était égale ou supérieure à 0,5 et l'entier immédiatement inférieur dans le cas contraire. Ce «reste» cumulé représente l'équivalent de 199 619,33 mots soit 0,5% de la surface théorique totale et 1,3% de la distance totale. Sans être considérable, l'incertitude n'est donc pas totalement négligeable ! Le même calcul, opéré sur un grand nombre de corpus, montre que ce défaut sera d'autant plus sensible que les vocables de basses fréquences occupent une surface importante (voir plus bas, la contribution à la distance calculée par classes de fréquences). Quand les textes comparés sont brefs (1 000 à 3 000 mots environ), la somme de ces décimales peut même dépasser 1,5% de la distance totale.

Le calcul et l'interprétation des indices doit donc respecter un certain nombre de précautions :

— pour limiter partiellement le poids de ces fantômes de mots, on n'appliquera pas le calcul à de trop petits textes (voir plus bas la discussion portant sur les classes de fréquence) : en tous cas, la taille du plus court doit dépasser 1 000 mots et on évitera une échelle des dimensions supérieure à 1/10 ;

— pour les mêmes raisons, nous éliminons du calcul tous les résultats inférieurs à 0,5 ;

Sous ces réserves, le problème des décimales oblige à inscrire les résultats du calcul des indices dans une marge d'incertitude de ± 1,5%. Par exemple, le plus petit indice observé dans le corpus Brunet sépare deux extraits du *Paysan parvenu* de Marivaux : 0,1952 (ou 1952 mots pour 10 000). Une présentation exacte indiquerait également les bornes de l'intervalle de confiance (0,1923 et 0,1981 mots). Dans le corpus Brunet, l'introduction de cette plage d'incertitude ne change rien, tant la structure par couples est évidente, mais ces précautions ne sont pas toujours inutiles, surtout dans des matières délicates comme l'attribution d'auteur…

2.2 L'influence de la taille

Ce problème peut être abordé à travers la propriété que nous nommons «transitivité» : quand on agrège le vocabulaire de deux textes, les indices des distances entre ce nouvel ensemble et les autres textes doivent refléter l'ordre des indices antérieurs.

L'enjeu de cette discussion est important. En effet, si les deux propriétés de l'indice — insensibilité à la taille des textes et transitivité — étaient vérifiées, l'indice de la distance entre le nouvel ensemble, constitué par l'agrégation de deux textes (B et C), par rapport à un troisième texte A, sera exactement égal à la moyenne arithmétique simple des deux indices séparant A de B et A de C :

$$(4)\ D_{a,(b+c)} = \frac{D_{a,b} + D_{a,c}}{2}$$

Cette propriété présenterait de nombreux avantages. En premier lieu, elle vérifierait l'insensibilité de l'indice aux variations de taille entre les textes comparés. Mais surtout, elle permettrait de simplifier les algorithmes de classification automatique, facilitant grandement la recherche de la meilleure partition possible au sein des vastes ensembles de textes…

En fait, les particularités du vocabulaire de toute langue naturelle interdisent a priori qu'une telle propriété puisse être exactement vérifiée (comme nous le verrons plus bas, cette question dépend fondamentalement de la déformation de la distribution des fréquences au fur et à mesure que la taille du corpus augmente). Mais l'approximation est-elle ou non acceptable ? à quelles conditions peut-on considérer que (4) est vraie ?

Nous allons d'abord esquisser une réponse empirique, en utilisant à nouveau le corpus Brunet dont certains couples de textes pouvaient être attribués sans hésitation à un même auteur. Nous avons fusionné ces textes et recalculé les indices des distances séparant ces nouveaux ensembles des 48 textes restants et nous avons comparé ce résultat avec ceux que donne la moyenne arithmétique simple ci-dessus. Comme ces fusions ont une taille approximativement double par rapport aux autres textes,

l'expérience renseigne à la fois sur l'influence des différences de taille et sur la «transitivité» de l'indice de la distance[4].

Dans un premier temps, nous avons utilisé le couple formé des deux textes les plus proches déjà mentionnés. Les différences entre les valeurs calculées grâce à (4) et les valeurs observées sont faibles, les écarts positifs équilibrant presque les écarts négatifs (la différence est de 1,4%). Autrement dit, sous réserve que les textes comparés soient relativement homogènes — comme l'est la fusion des extraits du *Paysan parvenu* —, les différences de taille semblent avoir une influence faible sur le calcul et la transitivité de l'indice se trouve vérifiée (sous réserve, là encore, d'une plage d'incertitude qui dépasse 3% dans un seul cas sur 48).

Nous avons réalisé la même expérience avec les deux textes séparés par l'indice le plus fort tout en étant probablement du même auteur (il s'agit des deux extraits tirés de : J. Verne, *De la terre à la lune*). Ils forment donc un couple hétérogène et, d'ailleurs, ils sont les plus décalés par rapport à tous les autres. Dans ce cas, les écarts, entre les valeurs observées et celles obtenues avec (4), sont un peu plus importants et généralement orientés dans le même sens : les valeurs observées sont assez systématiquement plus faibles que les valeurs attendues. La formule (4) surestime en moyenne les valeurs attendues de près de 3%, le biais atteignant même 5% dans deux cas.

Enfin, nous avons systématisé les deux expériences précédentes, en agrégeant, pas à pas, les textes dans l'ordre suggéré par la classification automatique, en observant, à chaque étape, les écarts entre les valeurs observées et celles calculées avec la formule (4). Le texte issu de cette fusion a une taille d'abord double par rapport aux autres, puis triple, etc.

On constate d'abord que la moyenne des observations a tendance à s'écarter progressivement des valeurs théoriques. Bien que l'accroissement de cet écart soit irrégulier, il est nettement corrélé avec l'augmentation des différences de taille et dépasse systématiquement 5%, lorsque la fusion dépasse cinq fois la dimension des autres textes. De plus, les écarts maximaux ont, eux aussi, tendance à s'accroître dans des proportions encore plus nettes.

---

[4] La taille et le format de la revue ne permettent pas de reproduire les tableaux de résultats. Ceux-ci peuvent être demandés à l'auteur.

Autrement dit, pour une analyse fine, il n'est pas inutile de garder en mémoire l'indice de la distance «intra» (celle qui sépare les textes fusionnés) afin d'en tenir compte dans l'interprétation des nouveaux indices des distances «inter» (séparant les fusions et les textes restants).

D'après les expériences menées jusqu'à présent, ces ±5% peuvent être considérés comme l'intervalle moyen d'incertitude dans lequel on s'inscrit :

— lorsque l'on compare des textes de dimensions assez différentes. Cet intervalle peut cependant être dépassé quand les différences de taille entre les textes sont trop importantes. En tout état de cause, il ne faut pas comparer des textes trop différents à la fois par leur vocabulaire et leur taille : 1 à 10 étant, de ce point de vue, un maximum à ne pas dépasser ;

— lorsque certains des textes comparés sont assez «hétérogènes», c'est-à-dire lorsqu'ils sont formés de plusieurs morceaux assez dissemblables ou qu'ils comportent une proportion inhabituelle de mots étrangers, de noms propres ou de termes techniques (cf plus bas) ;

— lorsque l'on considère comme vraie la propriété de transitivité de l'indice de la distance intertextuelle pour opérer des regroupements et des classifications. Cette incertitude sera d'autant plus forte que la taille du groupe formé par agrégation sera élevée par rapport aux autres textes ou groupes de textes…

Sous ces réserves, l'utilisation de la formule (4), dans les algorithmes de classification automatique ou d'analyse arborée, ne risque pas de conduire à des conclusions erronées. Et, en effet, les expériences de classification opérées avec les textes fusionnés du corpus Brunet donnent les mêmes résultats que ceux obtenus sur les fichiers d'origine : dans les arbres et les dendrogrammes, les fusions viennent se loger à la place des couples correspondants et le reste de la figure demeure inchangé. Par conséquent, l'utilisation de (4) ne modifie pas l'ordre de classement des textes, non plus que la «géographie d'ensemble» du corpus telle qu'elle ressort du degré relatif de proximité existant entre les différents groupes de textes. Or le problème essentiel d'une classification n'est pas le respect millimétrique des distances séparant chacun des individus classés mais la meilleure agrégation possible de leurs «ordres de préférence» respectifs. Cependant, il faut se souvenir que, dans les

graphiques récapitulant ces classifications, certaines «jambes» ou certains «troncs» seront probablement exagérés, que cette exagération sera négligeable lors des premiers regroupements, mais que le défaut sera plus sensible lorsqu'on approchera des agrégations ultimes.

Jusqu'à maintenant, ces intervalles de confiance et cette robustesse relative de l'indice se sont trouvés vérifiés dans la plupart des expériences que nous avons réalisées. Ils peuvent donc être utilisés comme étalons dans l'appréciation des résultats. Leur importance peut surprendre. Par exemple, lorsque l'on compare des textes de longueur différentes ou dont l'un au moins est assez hétérogène, la valeur de l'indice est comprise dans une plage d'incertitude de ±5%, ce qui revient à dire qu'un indice de 0,200 ne peut pas être considéré comme significativement différent d'un autre égal à 0,190 ou à 0,210…

3. La contribution à la distance

Au fond, on cherche à mesurer l'influence simultanée de l'auteur, de l'époque et du genre. En fait, on ne fait qu'observer des variations dans le vocabulaire qui est d'ailleurs hétérogène. On peut donc se demander quelles sont les parties du vocabulaire qui génèrent des ressemblances ou des dissemblances entre les textes. La mesure de la «contribution» des vocables à la distance apporte une réponse. En effet, en appliquant la formule (3), le programme commence par calculer, pour chaque vocable, la différence entre le nombre des occurrences effectivement observées dans un texte et la fréquence attendue dans un second. Pour chaque vocable (ou groupe de vocables), on compare cette différence avec l'indice calculé sur l'ensemble du vocabulaire des deux textes : si le résultat est supérieur à cette moyenne, le vocable contribue positivement à la distance (ou encore, il accentue le contraste entre les textes) ; à l'inverse les vocables, dont l'indice est inférieur à cette moyenne, y contribuent négativement (ils gomment les contrastes). Pour juger si ces écarts sont significatifs, on les rapporte à l'écart type calculé sur l'ensemble du corpus. Pour interpréter la dernière colonne des trois tableaux ci-dessous, il faut se souvenir qu'un écart inférieur à $\pm 1{,}98\sigma$ n'est pas significatif et que la probabilité de se tromper, en le considérant comme significatif, diminue au fur et à mesure que ce rapport augmente (au-dessus de $\pm 3\sigma$, les chances d'erreur sont inférieures à 1%).

Nous examinerons successivement la contribution à la distance des différentes catégories grammaticales, des vocables les plus fréquents et des classes de fréquence.

3.1 Les vocables classés en catégories grammaticales

Comme l'indique la dernière colonne du tableau I ci-dessous, les contributions, à la distance intertextuelle, des vocables classés en fonction de leurs catégories grammaticales, sont assez souvent significativement différentes de la moyenne. Il y a donc là une dimension caractéristique du langage qui n'est pas habituellement aperçue car la quasi-totalité des logiciels d'analyse textuelle travaillent sur les formes graphiques (considérées comme

des suites de lettres) sans référence à la langue et aux parties du discours.

Le groupe nominal est le principal facteur de différenciation entre les textes et, probablement, entre les époques comme l'avait déjà constaté E. Brunet dans son analyse du vocabulaire français de 1789 à nos jours (Brunet, 1981). Les noms propres viennent au premier rang, suivis des adjectifs et des substantifs. On vérifie ainsi le statut très singulier des noms propres dans la langue, considérée comme un code partagé entre les différents locuteurs : ils forment le point de contact privilégié entre ce code et la «réalité» extérieure (mais il s'agit ici de fiction). De plus, si l'on accepte le postulat de la théorie de l'information selon lequel un mot apporte d'autant plus d'information que son occurrence est moins prévisible, le substantif et l'adjectif pourraient bien former le cœur de la signification dans les textes analysés.

Tableau I Contribution à la distance des vocables classés en fonction de leur catégorie grammaticale

	Surface théorique	Distance absolue	Indice	Ecart réduit
Verbes	6 366 396	2 772 800	0,436	1,1
Noms propres	875 912	849 879	0,970	11,2
Substantifs	6 231 780	4 469 075	0,717	6,4
Adjectifs	1 874 175	1 379 615	0,736	6,8
Pronoms	3 610 946	1 163 505	0,322	-1,0
Autres pronoms	2 274 473	590 006	0,259	-2,2
Adverbes	2 879 919	1 035 453	0,360	-0,3
Déterminants	6 936 049	1 356 478	0,196	-3,4
Articles	4 717 431	598 724	0,127	-4,7
Adjectifs indéf.	377 097	108 744	0,288	-1,7
Prépositions	6 072 356	816 374	0,134	-4,6
Conjonctions	2 406 685	497 216	0,207	-3,2
Locutions, etc.	112 622	72 005	0,639	5,0

En revanche, les vocables appartenant au groupe verbal (verbes et pronoms) semblent globalement «banaux» : ils ne s'écartent pas significativement de la distance moyenne.

Cependant, cette moyenne forme une sorte de point d'équilibre entre le groupe nominal (dont la fluctuation est très forte) et la plupart des mots outils qui semblent plus stables. Il n'en reste pas moins que la plupart des verbes usuels semblent partagés par tous les auteurs et ne pas subir de variations significatives au moins dans ce corpus. Sans doute faudra-t-il examiner en détail ces verbes banaux. Cela permettrait de mieux comprendre le statut particulier du groupe verbal dans la langue.

Enfin, les mots-outils (adverbes, déterminants, prépositions et conjonctions) paraissent éminemment stables. Cela semble justifier les thèses selon lesquelles ces vocables sont largement communs à tous les usagers de la langue mais, pour autant, peuvent-ils être négligés lorsqu'on étudie les singularités de tel ou tel auteur ? En revanche, le calcul semble mettre en doute l'idée de certains statisticiens, notamment anglo-saxons, suivant laquelle les études d'attribution d'auteur doivent se concentrer sur ces mots outils parce que leur emploi serait largement automatique, et donc constant chez un auteur donné, quel que soit le genre dans lequel il s'exprime (Mosteller et Wallace, 1984). En effet, les résultats du calcul suggèrent que, au moins pour la plupart de ces vocables, les contrastes entre les auteurs sont assez faibles.

Cette discussion permet aussi de comprendre certains résultats de l'expérience menée avec E. Brunet. Par exemple, les textes les plus centraux sont de M. Proust, non pas que cet auteur ait, en quelque sorte, «synthétisé» tous les thèmes de la littérature mais, plus banalement, parce qu'il est celui qui, du fait de la structure complexe de ses phrases, utilise le plus de mots de liaison et de pronoms, présents dans tous les autres textes, mais avec des densités plus faibles, ce qui suffit à le placer au plus près du centre de gravité du corpus… A l'inverse, *De la terre à la lune* ou *Le cousin Pons* se caractérisent par la densité la plus élevée en noms propres, mots étrangers, substantifs et adjectifs, ce qui engendre une position «décalée» par rapport aux autres textes et une plus grande hétérogénéité.

L'examen des vocables les plus fréquents confirme ces intuitions.

3.2 Les vocables les plus fréquents

Pour mesurer la contribution des vocables les plus fréquents, on utilise le même raisonnement : l'écart entre la distance observée et la distance moyenne, mesurée sur l'ensemble du vocabulaire, n'est significatif que s'il excède ±1,98 σ. Ce calcul montre d'abord que les vocables les plus fréquents sont aussi les plus réguliers (l'indice moyen pour les 100 vocables les plus fréquents est à peine supérieur à 0,20). Cependant un petit nombre de vocables très fréquents sont aussi très irrégulièrement répartis (tableau II ci-dessous) : les pronoms de dialogue, les adjectifs possessifs, *monsieur*, *madame* et *cœur*. Au fond, le choix primordial pour un romancier se pose de la manière suivante : style direct ou indirect ; parler ou non d'amour[5]…

Tableau II Les vocables fréquents contribuant positivement à la distance entre textes

Vocable	Fréquence	Indice	Ecart réduit
madame	623	0,599	5,6
votre	592	0,592	5,5
tu	1190	0,591	5,5
coeur	463	0,527	3,9
vous	2560	0,517	3,5
monsieur	537	0,486	2,8
mon	2415	0,479	2,6
cela	544	0,477	2,4
moi	747	0,463	2,2
point (adv)	547	0,456	2,0
je	8755	0,455	2,0

A l'opposé, quelques vocables fluctuent très peu d'un auteur à l'autre. On peut ici limiter l'examen à la partie de cette population dont la répartition dans le corpus est si régulière que

[5] Une autre explication est possible : d'après l'étude d'E. Brunet sur le vocabulaire français, "monsieur", "madame" et "cœur" sont caractéristiques de la première moitié du XIXe et leur emploi recule considérablement par la suite (Brunet, 1981, III, p 77, 1238 et 1261).

leurs indices sont inférieurs à .20, moyenne observée pour les 100 vocables les plus fréquents (tableau III).

Ce dernier tableau ne surprendra pas. Il comporte les principaux outils de la langue française : les trois verbes les plus usuels (*être*, *avoir*, *faire*), les deux premières conjonctions de coordination (*mais*, *et*), les articles et prépositions les plus employés, le pronom relatif *qui*… Ces mots-outils, qui couvrent près de la moitié de la surface de tout texte, forment incontestablement le noyau central du code commun aux usagers d'une même langue.

Tableau III Les vocables fréquents les plus régulièrement répartis dans le corpus Brunet

Vocables	Fréquence	Indice	Ecart réduit
plus	2338	0,200	-3,3
il	10693	0,193	-3,5
se	4959	0,192	-3,5
mais	1726	0,192	-3,5
ce (article)	2938	0,182	-3,7
le (pronom)	4028	0,179	-3,7
tout (adj. indéf.)	1615	0,175	-3,8
faire	2547	0,174	-3,8
avoir	7829	0,161	-4,1
pour	2785	0,158	-4,1
qui	4158	0,154	-4,2
être (verbe)	9852	0,150	-4,3
en (prép)	3366	0,148	-4,3
avec	1666	0,143	-4,4
dans	3612	0,140	-4,5
le (article)	4028	0,129	-4,7
un (article)	9412	0,118	-4,9
et	10200	0,104	-5,1
de	30931	0,093	-5,3
à	11135	0,075	-5,7

En définitive, hormis les grands choix dont nous parlions plus haut, l'analyse des dissemblances oblige à descendre vers des éléments du lexique plus discrets et relativement moins visibles. On le comprend mieux en examinant la contribution des vocables rangés en classes de fréquences.

3.3 Les vocables classés par fréquences

Les vocables sont rangés en fonction de leurs fréquences dans l'ensemble du corpus : très basses (1 à 4), basses (5 à 9), etc. (tableau IV).

Tableau IV Contribution à la distance des vocables rangés par classe de fréquence

Classe de fréquence	% Distance totale	% Texte théorique	Indice de la distance	Ecart réduit
1 à 4	41,3	17,9	0,870	9,3
5 à 9	12,1	8,2	0,561	3,5
10 à 14	5,6	4,6	0,462	1,6
15 à 19	3,7	3,4	0,413	0,7
20 à 29	4,7	4,7	0,378	0,0
30 à 49	4,9	4,9	0,377	0,0
50 à 99	6,0	7,6	0,298	-1,5
100 à 199	6,3	10,4	0,229	-2,8
200 à 499	8,5	19,0	0,168	-3,9
500 à 999	1,9	2,6	0,269	-2,0
1000 et +	5,0	16,7	0,113	-5,0
Somme	100,0	100,0	0,377	

Ce classement permet de répondre à une question évidente : les ressemblances ou les différences proviennent-elles plutôt des vocables rares ou fréquents ?

Alors que les vocables très rares ne couvrent que moins du cinquième de la surface théorique, ils contribuent pour plus de 40% à la distance totale observée dans le corpus… A l'inverse, les quelques vocables très fréquents (tous verbes usuels ou mots-outils) ne contribuent que pour 5% à cette distance alors qu'ils occupent une surface presqu'équivalente.

Ces constats amènent des remarques importantes :

— la plupart des analyses — notamment ACP ou AFC — ne portent que sur les formes graphiques très fréquentes alors que ce sont manifestement les vocables «rares» qui sont le facteur principal de différenciation entre les locuteurs ;

— le problème des décimales, examiné ci-dessus, prend un relief particulier puisque ce sont avec ces vocables rares que les «fractions de mots» pèsent le plus lourd dans la distance ;

— toutes choses égales par ailleurs, des textes longs seront naturellement plus proches entre eux que des textes courts car les classes de fréquence élevées y pèseront proportionnellement plus lourd. En effet, les effectifs des classes de fréquences n'augmentent pas proportionnellement au fur et à mesure que le texte s'allonge. Certes, C. Muller a signalé la propriété selon laquelle les hapax (fréquence 1) seront toujours plus nombreux que les vocables survenant deux fois qui, eux-mêmes, excèderont les fréquences 3, etc. Mais, il n'en reste pas moins que, au fur et à mesure de l'allongement du texte, les vocables de basse fréquence occupent une proportion sans cesse plus petite de celui-ci et que, à l'inverse, le poids relatif des fréquences élevées augmente plus que proportionnellement. Or la formule (3) postule que la modification des fréquences est strictement proportionnelle puisqu'on applique le même coefficient de proportionnalité à tout vocable, quelle que soit sa fréquence. Cela milite donc pour qu'on limite le calcul à des textes de tailles pas trop différentes et ni trop petites ni trop grandes…

Pour surmonter cette limite évidente, plusieurs propositions ont été faites depuis notre article de décembre 2001. Ainsi, une équipe de Harvard a développé un calcul comparant non les fréquences des mots mais leurs rangs dans les distributions de fréquences de chaque texte (Yang et Al. 2003)[6]. D'autre part, T. Merriam préconise de prélever, dans tous les textes, des tranches de taille égale à celle du plus petit des textes comparés. Dans ce même numéro, il présente les résultats obtenus sur Shakespeare, en se limitant pour l'instant à un seul échantillon par texte (voir également Merriam, 2002). Il ne s'agit pas de prélever au hasard des mots sur toute l'étendue du plus grand texte — l'indice de la distance intertextuelle donne une bonne approximation du résultat de cette expérience — mais des *blocs* entiers. Naturellement, pour réduire la marge d'incertitude, il faut un nombre suffisamment grand de tranches différentes. Ces solutions sont en train d'être expérimentées et les résultats — pour l'instant prometteurs en ce

[6] Cette voie avait été suggérée par C. Muller à la fin du chapitre qu'il consacre à la "connexion lexicale" dans son manuel de 1977.

qui concerne la méthode préconisée par T. Merriam — feront l'objet de publications ultérieures.

Conclusions

L'indice de la distance intertextuelle fournit un bon outil pour la mesure des ressemblances et des dissemblances entre textes. Ses propriétés facilitent la recherche des meilleures partitions possibles, au sein de vastes bases de données textuelles, grâce à des techniques comme la classification hiérarchique ou l'analyse arborée.

Naturellement, comme dans toute estimation, les résultats sont affectés d'une certaine incertitude. Nous avons examiné successivement les effets des décimales, des différences de tailles entre les textes ou des variations de densité des catégories grammaticales… Bien sûr, ces effets ne se cumulent pas ; ils sont largement inclus les uns dans les autres. Par exemple, l'essentiel de l'«effet-taille» provient de ce que le poids des faibles classes de fréquence, donc des décimales, est plus important pour les petits textes que pour les grands, etc. Au total, on peut donc considérer que les mesures sont inscrites dans une plage moyenne d'incertitude de ±5%, autour de la valeur obtenue, à condition d'avoir respecté les précautions énumérées dans cette présentation. Loin d'être décevante, cette limite pourra paraître bien optimiste par rapport à la précision toute relative des mesures en sciences humaines et sociales…

Toutefois une dernière réserve mérite d'être émise. Les textes du corpus Brunet ont été passés au peigne fin : les fautes d'orthographe ont été corrigées puis les graphies ont été normalisées pour réduire à une seule forme les variantes dans l'écriture d'un même mot. Enfin, chacune de ces «formes normalisées» a été lemmatisée (c'est-à-dire qu'elle a été rattachée à son entrée de dictionnaire : vedette et catégorie grammaticale). Il faut porter un soin tout particulier aux majuscules et aux abréviations. Par exemple, en poésie, il est de tradition que chaque vers commence par une majuscule (soit un mot sur 6 à 8). Si l'on travaille avec les formes graphiques originales, sans réduire ces majuscules, on augmente mécaniquement l'indice de la distance entre la prose et la poésie d'au moins 13% ! De même, dans sa

correspondance, un auteur utilisera de nombreuses abréviations, acronymes, initiales de noms de personnes et autres raccourcis qu'il s'interdira certainement dans un ouvrage destiné à publication… Au total, classer des textes, qui auront été enfournés tels quels dans l'ordinateur, amènera probablement à conclure que les genres — confondus avec les conventions graphiques propres à chacun d'eux — sont plus importants que le thème, l'époque ou l'auteur.

La statistique commence à proposer des outils qui seront d'une grande utilité pour l'analyse des grandes bases de textes enregistrées sur support électronique. Mais cet intérêt nouveau soulève inévitablement le problème des normes d'enregistrement des textes et celui des conventions de mesure car il ne sert à rien de mesurer finement un phénomène dont la saisie n'aurait pas été assurée, au préalable, avec un minimum de rigueur.

Bibliographie

Baayen H. & Van Halteren H. & Tweedie F. (1996). «Outside the Cave of Shadows : Using Syntactic Annotation to Enhance Authorship Attribution», *Literary and Linguistic Computing*, 11-3 : 121-31.

Bergeron J.-G., Labbé D. (2000). «L'évaluation de la négociation raisonnée par les acteurs : une analyse lexicométrique», *XVIe congrès de l'Association Internationale des Sociologues de Langue Française*, Québec (à paraître aux Presses de l'Université Laval).

Binongo J.-N. & Smith M. W. A. (1999). «The Application of Principal Component Analysis to Stylometry», *Literary and Linguistic Computing*, 14-4 : 445-465.

Brugidou M. & Labbé D. (1999). *Le discours syndical français contemporain (CGT, CGT, FO en 1996-98),* Grenoble-Paris, CERAT-EDF(GRETS).

Brunet E. (1978). *Le vocabulaire de Jean Giraudoux*, Paris-Genève, Slatkine-Champion.

Brunet E. (1981). *Le vocabulaire français de 1789 à nos jours*, Paris-Genève, Slatkine-Champion.

Brunet E. (1988a). «Une mesure de la distance intertextuelle : la connexion lexicale», *Le nombre et le texte. Revue informatique et statistique dans les sciences humaines*, 24, Université de Liège.

Brunet E. & Muller C. (1988b). «La statistique résout-elle les problèmes d'attribution ?», *Strumenti critici*, III, 3 : 367-387.

Gower J.-C. (1985). «Measures of similarity, dissimilarity and distance», in Kotz S., Johnson N.-L., et Read C.-B. (eds), *Encyclopedia of Statistical Sciences*, Vol 5, New York, Wiley : 397-405.

Holmes D. (1995). «The Federalist revisited : new directions in autorship attribution», *Literary and Linguistic Computing*, 10-2 : 111-127.

Hubalek Z. (1982). «Coefficients of Association and Similarity, based on Binary (Presence Absence) Data : an Evaluation», *Biol. Rev.*, 57 : 669-689.

Jaccart P. (1908). «Nouvelles recherches sur la distribution florale», *Bull. Soc. Vaud. Sci. Nat.*, 44.

Labbé C. & Labbé D. (2001). «Inter-Textual Distance and Authorship Attribution Corneille and Moliere», *Journal of Quantitative Linguistics*, 8-3, December 2001, p. 213-231.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahiers du CERAT.

Labbé D. & Hubert P. (1998). «La connexion des vocabulaires», in Mellet S. (ed*), 4e Journées d'analyse des données textuelles,* Nice : 361-370.

Labbé D. & Monière D. (2000). « La connexion intertextuelle. Application au discours gouvernemental québécois », Rajman M. & Chappelier J.-C. (eds). *Actes des 5e journées internationales d'analyse des données textuelles*, Lausanne, Ecole polytechnique fédérale, vol 1 : 85-94.

Merriam T. (2002). "Intertextual Distances Between Shakespeare Plays, with Special Reference to *Henry V* (verse)". *Journal of Quantitative Linguistics,* 9-3, December 2002 : 261 - 273.

Mosteller F. & Wallace D. L. (1984). *Applied Bayesian and Classical Inference : The Case of the Federalist Papers*, Addison-Wesley, Reading.

Muller C. (1977). *Principes et méthodes de statistique lexicale*, Paris, Hachette université.

Rudman J. (1998). «The State of Authorship Attribution Studies : Some Problems and Solutions», *Computers and the Humanities*, 31 : 351-365.

Yang A. C.-C. & Al. (2003). "Information Categorization Approach to Literary Authorship Disputes". A paraître dans *Physica A*, 2003.